Usability Metrics for Spoken Language Systems

Priyanka Gupta  
*Auburn University*  
107 Dunstan Hall  
*Dept. of Computer Science & Software Engineering*  
*Auburn University, AL, 36849 USA*  
guptapr@auburn.edu

Juan E. Gilbert  
*Auburn University*  
107 Dunstan Hall  
*Dept. of Computer Science & Software Engineering*  
*Auburn University, AL, 36849 USA*  
gilbert@auburn.edu

**Abstract.** This paper explores an existing spoken language system evaluation method and presents an alternative method, called the “Speech Usability Metric”. The limitations of the pre-existing method are explored and a much simpler approach to the evaluation of spoken language systems is presented. The Speech Usability Metric calculates the extent to which a system is usable based upon the designer’s intent of how the interface will be used.

**Keywords:** Interface Evaluation, Spoken Language Systems, Speech Usability Metric
1. Introduction

As Spoken Language Systems are being used more often for a wide range of applications, there is an increasing demand for standardized benchmarks to test and compare performance and usability. Most usability studies do not evaluate the system beyond speech recognition accuracy, adequate language modeling, and appropriate semantic representations for efficient interaction with back-end knowledge sources [Walker, 2000]. Significant progress has been made towards identifying standards to achieve this goal. A wide variety of measures have been used, including measures like task success. Other metrics evaluate user satisfaction in conjunction with task success.

Paradise (PARAdigm for Dialogue System Evaluation) is one such method for evaluating spoken dialogue systems [Walker, Litman, Kamm, Abella, 1997]. Using Paradise, it is possible to compare the performance of different dialogue strategies, to evaluate performance. It is also possible to compare systems designed for different tasks. This method is one of the most recognized for evaluating spoken dialogue systems, as it uses user satisfaction ratings as an indicator of usability. It calculates the usability as a measure of two factors: task success and dialogue costs using a decision-theoretic framework. The Paradise framework requires a complex and data collection intensive process. Furthermore, the Paradise framework aims to maximize user satisfaction by maximizing task success and minimizing costs. This is an appropriate approach for most systems, but there are cases where user satisfaction is less important than task completion time, e.g. military applications. The goal of this research was to design an alternative usability metric that addresses the shortcoming of the Paradise framework.

A Speech Usability Metric (SUM) was developed and tested using two Spoken Language Systems. This evaluating method tests Spoken Language Systems on using three metrics, namely user satisfaction, accuracy and task completion time. However, it is flexible enough to test along additional metrics. User satisfaction was calculated using questionnaires. Interaction between the user and the system was recorded to calculate the remaining two metrics. The development of any new method or interface is typically an iterative process. This paper represents the first iteration of the development of a SUM. Over time this method can be refined to fit the needs of the designer.
2. Literature Review

2.1 Evaluating User Interfaces

“Every designer wants to build a high-quality interactive system that is admired by colleagues, celebrated by users, circulated widely, and imitated frequently” [Shneiderman, 1997].

A necessary aspect of designing such an interface is evaluation. Evaluating user interfaces is becoming one of the most important aspects of software development. There are many factors that determine whether an interface is good or bad.

Some of them are as follows:

- Functionality
- Speed and efficiency
- Reliability, security, data integrity
- Standardization, consistency
- Usability

In evaluating an interface, the most significant criterion is the usability of the interface. Interactive applications require the end users to comprehend the interface, so that they may complete the tasks at hand. An interface will be inherently usable if all phases of the design process are user-centered. The system should be designed so that it is tailored to the needs and requirements of the user and not the other way around [Walker, 2000]. However, there are applications that are mission critical that do not emphasize user satisfaction. These interfaces require training and typically have minimum performance requirements.

There are several methods used to evaluate the usability of an interface. Formative evaluation is one of them. It requires evaluating a user interface during all stages of development. It is an integral part of prototyping the system. It involves the collection of subjective, objective, quantitative and qualitative data, which is later analyzed. The results are used to fine tune the interface in as part of an iterative cycle.

Graphical user interfaces have a well-defined set of methods that can be used to make an interface more user-friendly. Some of the major advances in this area are made using style guides. Systematic testing involving actual users is still one of the most reliable methodologies for developing user-friendly software
systems. However, the same evaluation techniques cannot be applied to spoken language systems. Evaluating a spoken language system requires a different set of considerations.

2.2 Evaluating Spoken Language Systems

“A spoken language system by definition is a system which relies primarily or exclusively on audio for interaction, this includes speech and sound [Kamm, Walker, 1997]. Speech is hailed to be one the most natural modes of interaction. Thus, the numbers of speech-enabled applications are rapidly increasing. Enterprises, government agencies, and wireless carriers are all choosing voice over touchtone for their automated telephone applications. In order to effectively design strong spoken language systems, designers need equally robust methods to test and evaluate them. Users want and expect to be able to speak and be understood with no undue effort when they use these auditory interfaces.

There are well-developed sets of methods that are used to evaluate graphical user interfaces. A subset of these same principles can be applied to spoken language systems, as they are equally relevant for creating usable auditory interfaces. A spoken language system needs to be designed so that the user understands the capabilities and limitations of the system. The system should guide the users to utter predictable phrases to avoid unconstrained dialogue, just as we do when interacting with another human.

Human short term memory is limited to seven plus or minus two items [Miller, 1956]. This is a very important factor in spoken language systems design. Graphical user interfaces are persistent over time. When a user is viewing a graphical user interface, s/he can page down or walk away from the interface and the data will persist on the screen. Spoken language is transient; once something is said, it is gone. Therefore, all data must be maintained in the user’s short term memory. This is a major difference between graphical user interfaces and auditory user interfaces. Auditory user interfaces use speech recognition to accept input from users. Speech recognition is imperfect. This imperfection often results in higher error rates versus user input on graphical user interfaces. Therefore, error recovery becomes an essential aspect of all spoken language systems.

Error recovery has to be built into the system to the extent that users can gracefully recover from speech recognition errors. However, this is for designers to know, not for users to comprehend. Misrecognitions, time-outs and out of vocabulary recognition errors are all aspects of every spoken language system that must be addressed during design.
All of the above mentioned requirements are incorporated into the Paradise Framework [Walker, Litman, Kamm, Abella, 1997]. Though the Paradise model addresses the issues concerned with evaluating spoken language systems, it is a very complex method and it does not address systems that do not emphasize user satisfaction.

The primary focus of this research is to investigate an alternative approach to the Paradise framework, called the Speech Usability Metric (SUM). Therefore, it is essential to review them first. The following section discusses the Paradise Framework in a greater detail.

2.3 Paradise Framework

2.3.1 Introduction

Even the interfaces that are designed using all the principles of a good user interface design cannot guarantee a successful spoken language system. Spoken language systems are comprised of many different interwoven components. The success of the spoken language system depends on whether the user was able to accomplish the task successfully using the system and how the task was actually accomplished. This implies that an extremely critical step in the development of a spoken language system is not only the designing of the interface and devising dialogue strategies but also evaluating how the system performs. Performance of the system includes inappropriate utterance ratio, turn correction ratio, concept accuracy, implicit recovery, number of turns, elapsed time and other such factors which make the whole process of comparing dialogue strategies across system and tasks difficult.

Paradise (PARAdigm for DIalogue System Evaluation) is a general framework used for evaluating spoken dialogue agents or the spoken language systems [Walker, Litman, Kamm, Abella, 1997]. It encompasses and enhances all the previous research conducted in this area. It helps in evaluating different strategies for the same task. In order to achieve this, it separates the task from the strategies used by the dialogue agent to achieve the task goals. Paradise framework defines performance quantitatively as a weighted function of a task-based success measure and dialogue-based cost measures, where weights are computed by correlating user satisfaction with performance [Walker, 2000]. This helps to determine the tradeoffs among various factors that contribute to performance. It also allows performance to be calculated from sub-dialogues as well as whole dialogues. The following sub-sections will briefly discuss the Paradise’s framework and the way it is applied.
2.3.2 Design and Calculations

The Paradise framework uses methods from decision theory to combine a discordant set of performance metrics. For example, user satisfaction, task success and other such variables are plugged into a single evaluation function. The use of this theory requires a listing of both objectives of the decision problem and a set of measures for operationalizing the objectives [Walker, Litman, Kamm, Abella, 1997].

Paradise is based on a structure of objectives, which is shown in Figure 1.

![Figure 1: Structure of Objectives for Spoken Dialog Performance](image)

The Paradise model states that performance can be meaningfully correlated to values such as usability. Thus, the overall goal of a spoken language system is to maximize an objective that is related to the usability of the system, which can be measured directly as user satisfaction. The use of user satisfaction as a criterion to calculate system performance is based on the assumption that user satisfaction predicts other objectives such as willingness to use or pay for a service which, in themselves, are not easily measured. Two other factors that are used by this framework are task success and dialogue costs as contributors towards user satisfaction. Linear regression is used to quantify the relative contribution of success and cost factors to user satisfaction. The framework makes use of an attribute value matrix (AVM), which is used to represent what the system wishes to accomplish; that is, what information is required by the system and what information must be conveyed to the user in order to successfully complete the task.
Table 1 shows an example of an attribute value matrix for a simple information access dialogue in the train timetable domain.

<table>
<thead>
<tr>
<th>ATTRIBUTE</th>
<th>POSSIBLE VALUES</th>
<th>INFORMATION FLOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depart city (DC)</td>
<td>Milano, Roma, Torino, Trento</td>
<td>To agent</td>
</tr>
<tr>
<td>Arrival-city (AC)</td>
<td>Milano, Roma, Torino, Trento</td>
<td>To agent</td>
</tr>
<tr>
<td>Depart range (DR)</td>
<td>Morning, Evening</td>
<td>To agent</td>
</tr>
<tr>
<td>Depart time (DT)</td>
<td>6am, 8am, 6pm, 8pm</td>
<td>To user</td>
</tr>
<tr>
<td>Request type (RT)</td>
<td>Reserve, purchase</td>
<td>To agent</td>
</tr>
</tbody>
</table>

Table 1: Attribute value matrix, train timetable domain with requests

Task success is calculated by examining how well the system and the user achieve the information requirements of the task by the end of each dialogue, as measured by the Kappa coefficient. The Kappa coefficient is a measure of the relationship between the observed AVM and the ideal AVM and also of the results attributable to chance.

System performance is calculated by combining task success with cost measures. Intuitively, cost measures are calculated based on users or dialogue behaviors that should be minimized. They include efficiency measures and qualitative measures.

Therefore, evaluating a spoken language system using the Paradise model involves having a group of users perform tasks with known “ideal” outcomes, measuring user satisfaction, and measuring cost measures for the user-system dialogues.

The Speech Usability Metric is an alternative approach to the Paradise framework. Two spoken language systems were used to evaluate the Speech Usability Metric. These two systems are described in the sections that follow.

3. System Design and Description

3.1 Introduction

Since this research involves evaluating the Speech Usability Metric (SUM), two very different spoken language systems were used. One was a spelling system and the second was the professionally constructed HeyAnita system. Both system overviews are given in the subsequent sections in greater detail.

3.2 Spelling System Overview

A telephony based spelling system for words was constructed for this study. The spelling system was constructed to provide a system with likely recognition difficulties to contrast against a professionally
developed system, HeyAnita. Users of the spelling system can spell or pronounce a word or phrase. The system tries to catch the word. If unable to do so, the system asks the user to spell the word. The system checks with the user after each letter to make sure that it recognized the correct letter. If the user’s answer is negative, then the system tries to recognize the correct letter by pronouncing similar sounding alphabet letters. For example, if the user spells out the word cat; and the system recognizes “t” incorrectly, then it goes through all the sound alike letters “e, c, t, g, d etc.” to get the user to input the correct letter. All user input is prompted with a system beep. This beep helps the user to determine the time that should elapse between two utterances and makes it easier for the system to recognize user input. Once the user is able to input the whole word correctly, the system repeats the word and the user hangs up.

3.3 HeyAnita System Overview

HeyAnita Incorporated [HeyAnita.com] provides enhanced-communication software products and solutions to the telecommunications industry. They provide software solutions to technical problems as well as hosting and professional services. For this study the HeyAnita voice portal was tested for a specific task. The voice portal provides the user with a guide to relevant information over the phone. In order to access the system the user simply has to call 1-800-44-ANITA and say his or her choice from the keywords provided. A user can register to personalize his or her choices. Some of the choices offered by the HeyAnita system are:

1. Main Menu
2. Anita Express: Short cut to pre-selected information.
3. Flight Tracker: To check the status of domestic as well as international flights that depart from and arrive at most U.S. airports.
4. Horoscope: User can select horoscope by specifying zodiac sign or date of birth.
5. Lottery: Lotto, Powerball and other lottery results, categorized by states.
6. Newsroom: All sorts of different news ranging from world news to entertainment news is available.
7. Sports: Myriad information on various popular sports available.
8. Stock Portfolio: Over 15,000 quoted available for stocks traded on the NYSE, NASDAQ and AMEX.

9. TV Dish: Synopses for selected prime time serials, soap operas and other popular television shows.

10. Weather: Five-day forecasts for weather in over 3,000 U.S. and 300 international cities.

As is the case for most voice applications the user is mobile and informed because it can be accessed from anywhere at any time. HeyAnita’s voice browsing applications leverage content and information from various third parties. Information access and update for most of these content sources is achieved via XML integration and FTP transfer.

HeyAnita’s voice browsing applications are supported by its own FreeSpeech Gateway Server. It is a scalable software solution responsible for speech processing and call control capabilities. The FreeSpeech platform powers the voice browser (full VoiceXML support) and other applications with complete caching and grammar management. The HeyAnita voice portal was selected because of its easy to use interface versus the more complicated spelling system’s interface. Next, the usability metric used to evaluate the tasks given to the user using the above two systems, the spelling system and the HeyAnita system [HeyAnita.com] will be discussed.

4. Speech Usability Metric

4.1 Introduction

Evaluation is a prerequisite to designing effective and natural interfaces. The methods used to evaluate graphical user interfaces are well established and numerous. In order to design effective and natural spoken language systems, an iterative process is used. This includes cycles of design, implementation, experimentation with users, and evaluation, followed by redesign and implementation of improvements, based on the results of the user tests. When conducting this iterative process, when is the speech interface good enough? What metrics determine success? Without answers to these questions, this iterative process could go on indefinitely and become very costly. Therefore, the evaluator must have some concept of when the interface is good enough to end user testing. The Speech Usability Metric (SUM) was
designed such that the designer can specify metrics that are relevant to the interface and specify metric goals that will determine when the interface is good enough. The SUM is stated as follows:

\[
\text{SPEECH USABILITY METRIC} = X \times \text{(USER SATISFACTION)} + Y \times \text{(ACCURACY)} + Z \times \text{(TASK COMPLETION TIME)} \text{ Where } X + Y + Z > 0 \text{ and } X, Y, Z > 0.
\]

4.2 Explanation

The SUM uses three weighted metrics: User Satisfaction, Accuracy and Task Completion Time. The weights are specified by the designer with respect to each metric’s importance. For example, User Satisfaction may be weighted at 20%, Accuracy at 30% and Task Completion Time at 50% for a mission critical military application. The Paradise Framework doesn’t have this flexibility where the designer can specify metrics and weight them based upon their importance relevant to the designed interface. As described above, the SUM uses User Satisfaction, Accuracy and Task Completion Time. However, these metrics are not written in stone. These metrics were defined for the experiments described in this article. The designer could specify other metrics and assign weighted values as necessary.

User Satisfaction is typically the most heavily weighted of the three metrics. Typically, User Satisfaction is measured using Surveys and/or Interviews on a Likert scale. Accuracy measures the effectiveness of the speech recognition engine. Speech recognition errors can be categorized as misinterpretations (the system recognized the wrong word(s) or false positives), out of vocabulary errors (system didn’t recognize what the user said) and incorrect choices (the user says the correct word(s), but the user chose the wrong path). Accuracy can be measured by counting the number of recognition errors for each user. Task Completion Time (TCT) measures how long it takes the user to complete a task or dialogue. TCT is can be measured by timing the users conversation from beginning to the task completion. Additionally, TCT can be measured as follows:

- Establish a Maximum Task Completion Time (ETCT), which may be the largest acceptable TCT and compare all users’ TCT against the MTCT.
- Allow experts to use the system and those experts can then establish an Expected Task Completion Time (ETCT). Through experimentation, compare all users’ Task Completion Time to the Expected Task Completion Time.
o If there are no experts, calculate the Task Completion Time of each user and develop an average and standard deviation for Expected Task Completion Time based upon the users. Once this is established, a calculation of the Task Completion Time percentage relative to the Expected Task Completion Time can be made.

The SUM includes three weights, X, Y and Z. These three weights have values between 0 and 1 and their sum must equal 1; \( X + Y + Z = 1 \). They are set to a value defined by the client or the designer. These values are system dependent. The value of these weights depends upon the type of speech system to be evaluated. A safety critical system may have a high accuracy weight requirement, e.g. 0.80, versus a customer service system where user satisfaction is more important.

5. Experiment and Analysis

5.1 Introduction

An experiment was conducted to evaluate the effectiveness and usability of the spelling system versus the HeyAnita system using the SUM. The goal of this study was to take two very different systems and compare them using the SUM in an effort to determine the degree to which the two systems differ on usability. The professionally developed HeyAnita system was selected because it was hypothesized to be a more usable system versus the spelling system. The SUM was used to measure the extent to which the HeyAnita voice portal is more usable than the spelling system.

5.2 Experiment Protocol

5.2.1 Participants and Procedures

Forty-five Computer Science graduate and senior undergraduate students participated in this experiment at Auburn University. There were seventeen female and twenty-eight male participants. Three of the forty-five students were outreach students (They did not attend the lectures in person, but viewed the videotapes of the class throughout the semester). The experimental procedure was kept exactly the same for all the students whether they were outreach students or full time students. Any deviations in the protocol are noted.
Participants were informed that the purpose of this experiment was to complete the given task by retrieving the results to a given question. As both spoken language systems are different, the tasks allotted to the student for each system were also different. In the HeyAnita voice portal, students were asked to obtain the weather in the city of Auburn, Alabama at the time during which the experiment was conducted. In the spelling system, the students were asked to spell the word “JUDY”. If the system recognizes “JUDY”, it will spell it back to the student, which is noted as successful completion of the task. Accordingly, they were given a list of instructions outlining these tasks. These instructions specified the target tasks for each system and the protocol for the experiment. Participants were also informed that their interaction with the system would be recorded, so that they could be analyzed at a future date for accuracy and task completion time measures. All the participants were informed that the experimenter was there as an observer and would not provide any assistance that involved the specified task. Every participant was asked to complete the given task in each of the two systems. To avoid user familiarity after having used the first system, participants were asked to test the system in alternate order. For instance, if participant one is asked to test the HeyAnita voice portal first, followed by the spelling system, the next participant was asked to test the spelling system first followed by the HeyAnita voice portal.

Each participant worked alone. Each participant used the same PC and the same phone. The experimenter was present in the room but there was no interaction between the experimenter and participant while the experiment was conducted. The participants were asked to initiate the phone session without assistance from the experimenter. At the end of the experiment, participants were given two questionnaires, one for each system. The content on each questionnaire was identical. However, the first page of the questionnaire collected demographic information about the individual participants. Upon completion of the questionnaire, students were instructed not to talk to their classmates about the experiment, to ensure that all the participants had an equal amount of knowledge of the experiment prior to participation.

5.2.2 Materials

There were two phones used for this experiment. One was connected to a computer to record the conversation between the participant and the system and the second line was used by the participant to
connect to the HeyAnita voice portal or the spelling system. For the outreach students, a three-way call was made. The experimenter initiated this three-way call. Before the actual data and deductions are discussed, a look at the concepts used to analyze the data is necessary.

5.3 Analysis Discussion

In order to evaluate the effectiveness of the Speech Usability Metric (SUM), two systems were compared that are hypothesized to be very different with respect to usability. There are three steps that outline the methodology used in preparation for the analysis of this experiment:

1. Definition of a task;
2. Calculation of user satisfaction using surveys;
3. Calculation of accuracy values and task completion time using quantitative measures.

Conclusions were derived using the SUM for spoken dialogue evaluation. Most of the questions in the questionnaire were answered by the participants using a Likert-like scale. This quantifies the data and makes it easy to analyze. The questionnaire fully provides the requisite data for calculating User Satisfaction. User Satisfaction is calculated by summing the degree of user-agreement on a nine point Likert-like scale. For each question, nine points represent the best performance; one-point represents the worst performance. The participants filled out surveys that indicated user satisfaction and user perception of task completion. To assist in the determination of task success, the recorded interaction between the user and the system was used. The other measures were calculated as described below.

For accuracy and task completion time, an expert user was asked to complete the tasks. Since the expert user was already familiar with the system, the number of prompts he made to complete each task in both systems showed a significant and consistent decrease. Subsequently, the number of utterances that each participant made was recorded. Calculations were made to further analyze and average out the data in order to determine the correct value for accuracy with respect to the expert’s accuracy scores. Task completion time is calculated by writing down the exact amount of time it took each student to complete the task and then comparing it to the expert’s task completion time. In order to use user satisfaction scores, accuracy scores and task completion times in the SUM, the values for accuracy and task completion time had to be scaled to a nine point scale. This guarantees that all data points are on the same number scale.
To calculate the quantitative data for user satisfaction, accuracy and task completion time, the following equations were used:

- **User Satisfaction** = The average value obtained from each user satisfaction question from the survey.
- **Accuracy** = The absolute value of \((\text{Number of Prompts made by the Participants} - \text{Number of Prompts made by the expert}) / \text{Number of Prompts made by the expert}\).
  
  When a recognition error occurs, the user is given an additional prompt that attempts to guide the user to an appropriate response. Therefore, the number of prompts was used to measure speech recognition accuracy. As the number of prompts given to a participant increased, the value of accuracy decreased. Taking this into account, the value of accuracy was scaled to a nine point scale. As mentioned before the scaling was done to ensure that user satisfaction, accuracy and task completion time are operating on the same numerical scales. This was done by taking the highest value obtained using the accuracy equation for each system and then dividing that value by a factor of nine. For example, in the spelling system the highest number of prompts made by a participant was 38; using the accuracy equation, the accuracy value obtained was 11.67. Approximating this value to 12 and then dividing this value by 9, yields a scaled accuracy of 8.75.

- **Task Completion Time** = The total time taken to complete a task in seconds.
  
  The task completion time was scaled similarly to the accuracy because of the relationship between task completion time and the nine point Likert-like scale. For example, as the task completion time increases, the appropriate value on the Likert-like scale should decrease. Taking this into account, all task completion times were scaled using similar calculation used to scale accuracy.

5.4 Results and Discussion

The overall Speech Usability Metric was calculated from both sets of data collected, one for each system as discussed in the section above. In this section, the results from these two sets are analyzed. The
Speech Usability Metric given below was used to measure and compare the usability between the spelling system and the HeyAnita voice portal.

SPEECH USABILITY METRIC = X * (USER SATISFACTION) + Y * (ACCURACY) + Z * (TASK COMPLETION TIME)

With the value of weights as follows:
- X = 0.80, 0.70, 0.60
- Y = 0.15, 0.20, 0.25
- Z = 0.05, 0.10, 0.15

The results found for these data are analyzed and discussed in the sections below.

5.4.1 User Satisfaction Analysis

From the collected data for the first set of metrics, several findings were discovered.

1) Out of 45 students who participated in the study, only five answered affirmatively that the spelling system was better than the HeyAnita system. The Speech Usability Metric also supported this finding that the HeyAnita system was better than the spelling system.

2) The value for Speech Usability Metric for the HeyAnita system was higher than that for the spelling system. The overall values for user satisfaction, accuracy and task completion time were significantly higher for the HeyAnita system versus the spelling system.
3) Figure 2 contains the average user satisfaction ratings for each question on the experiment questionnaire. The participants were given thirteen different questions related to the usability of each system. The average user satisfaction across all questions for the HeyAnita voice portal was 83%. Likewise, the overall user satisfaction rating for the spelling system was 53%, figure 4. These figures clearly illustrate that the HeyAnita voice portal is much more user friendly than the spelling system.

4) Under the spelling system approximately 43% of the participants were unable to complete the task. However, in the HeyAnita voice portal, only 4% (2 students) of the participants were unable to complete the task. The results collected using the questionnaire consistently supported that the HeyAnita voice portal was superior to the spelling system when it came to user satisfaction.
5) Calculation of user satisfaction of an auditory interface also depends on the extent to which the system is predictable. This was included as a question on the questionnaire. Figure 3 depicts the chart describing the comparison between the predictability of the HeyAnita voice portal and the spelling system. According to the participants’ responses, approximately 88% of the users found it easy to predict the behavior of the HeyAnita system. None of the users had ever used a similar system before, yet they did not have problems with performing the task assigned. However, 57% of the users had problems with predicting the system behavior in the spelling system.
5.4.2 Accuracy Data Analysis

The second set of metrics collected were related to accuracy.

Table 2: The Descriptive Statistics of Accuracy

<table>
<thead>
<tr>
<th></th>
<th>MEAN</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPELLING SYSTEM</td>
<td>6.96</td>
<td>26</td>
</tr>
<tr>
<td>HEYANITA SYSTEM</td>
<td>8.66</td>
<td>43</td>
</tr>
</tbody>
</table>

1) Table 2 shows the accuracy statistics across the systems. The average value of accuracy for the spelling system after scaling the values into a Likert-like scale was 6.96 and, for the HeyAnita voice portal the accuracy value was 8.66. Table 2 also has a column titled N, which refers to the number of students who were able to complete the assigned tasks for each system. Therefore, accuracy values were measured for participants that successfully completed the tasks. Notice that only 26 participants were able to complete the assigned tasks using the spelling system versus 43 for the HeyAnita voice portal. Recall, that accuracy was measured by comparing the number of prompts made by each participant and then comparing that to the number of prompts made by the expert. For both systems, the number of prompts made by the expert was 3. For example, if the number of prompts made by a participant was 5 in the HeyAnita voice portal, then the accuracy value from the equation used would be taking the absolute value of [(3-5)/5], which
yields 0.67. Next, the value 0.67 is scaled onto the Likert-like nine point scale. Finally, the accuracy value for this participant is obtained by subtracting 0.67 from 9, which yields an accuracy value of 8.5.

The accuracy values for both systems were not significantly different. However, the failure rate for the HeyAnita voice portal was 4% versus 42% for the spelling system. In other words, 4% of the participants could not complete the HeyAnita voice portal task versus 42% for the spelling system. This statistic points to the fact that, though the average accuracy for the participants that successfully completed the tasks was not significantly different; the HeyAnita voice portal had a much higher success rate.

![Spelling System Prompts](image)

**Figure 5:** Actual Prompts made by the users in the Spelling system
2) Figure 5 shows the actual number of prompts made by each user of the spelling system. The actual number of prompts required by an expert was 3. None of the experiment participants reached this mark. The overall average number of prompts made for the spelling system was 11. Figure 6 shows the same statistics for the HeyAnita voice portal. Approximately, 40% of the students were able to meet complete the task using 3 prompts, which matches the expert.

5.4.3 Task Completion Time Data Analysis

The task completion time for both systems was about the same. The task completion time for the HeyAnita voice portal could have been higher, but the system enumerates all of the menu options to the user upfront. Given all the participants were first time users of the HeyAnita voice portal, they all listened to these prompts, which increased their task completion time. After experiencing the HeyAnita voice portal, a return caller could barge-in during the prompts and significantly reduce the task completion time. The spelling system only does one task, so the user does not have to wait for it to list several choices.
The task completion times for the expert were 48 seconds and 60 seconds on the spelling system and the HeyAnita voice portal, respectively. The expert took 13 seconds less to complete the task in the spelling system versus the HeyAnita voice portal; therefore, it was expected that the participants would also take less time to complete the task in the spelling system versus the HeyAnita voice portal. However, the average task completion times were 143 seconds and 140 seconds for the spelling system and the HeyAnita voice portal, respectively. The maximum amount of time taken to complete the task was 325 seconds for the spelling system versus 264 seconds for the HeyAnita voice portal. These contradictions between the expert and the participants can be explained by the amount of time students took to recover from errors. Students took much more time to recover from their errors in the spelling system versus the HeyAnita voice portal. The system design of HeyAnita facilitated fast recovery; hence, on average, it had better task completion times.

**Figure 7**: Task Completion Time for the Spelling System

Figures 7 and 8, show the task completion time for all the students that participated in the study.
5.4.4 Speech Usability Metric (SUM)

Table 3 describes the calculations of Speech Usability Metric for these experiments.

SPEECH USABILITY METRIC = X \times (USER SATISFACTION) + Y \times (ACCURACY) + Z \times (TASK COMPLETION TIME) User Satisfaction

- User Satisfaction = U_S
- Accuracy = A
- Task Completion Time = TCT
- Speech Usability Metric = SUM

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>U_S</th>
<th>Y</th>
<th>A</th>
<th>Z</th>
<th>TCT</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPELLING SYSTEM</td>
<td>0.8</td>
<td>4.9</td>
<td>0.15</td>
<td>6.96</td>
<td>0.05</td>
<td>5.15</td>
<td>5.222</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>4.9</td>
<td>0.20</td>
<td>6.96</td>
<td>0.10</td>
<td>5.15</td>
<td>5.358</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>4.9</td>
<td>0.25</td>
<td>6.96</td>
<td>0.15</td>
<td>5.15</td>
<td>5.4525</td>
</tr>
<tr>
<td>HEYANITA SYSTEM</td>
<td>0.8</td>
<td>7.46</td>
<td>0.15</td>
<td>8.66</td>
<td>0.05</td>
<td>5.36</td>
<td>7.535</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>7.46</td>
<td>0.20</td>
<td>8.66</td>
<td>0.10</td>
<td>5.36</td>
<td>7.49</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>7.46</td>
<td>0.25</td>
<td>8.66</td>
<td>0.15</td>
<td>5.36</td>
<td>7.445</td>
</tr>
</tbody>
</table>

Table 3: Calculation of SUM
Based on the calculations shown in Table 3, it is clear that the HeyAnita voice portal is more usable than the spelling system as determined by the SUM. Table 3 shows that the HeyAnita voice portal’s SUM is better than the spelling system’s SUM across all the defined variations of the usability weights. A perfect SUM value for these experiments would be 9. Table 3 shows the HeyAnita system is 1.5 points away from a perfect SUM score. Using these results, a speech user interface designer can see the weak areas for both systems. Furthermore, the designer has a final SUM that determines the extent to which these systems are usable based upon the predefined metric weights.

6. Conclusions

A Speech Usability Metric (SUM) was developed and tested using a comparative analysis between two speech user interfaces in an effort to offer an alternative approach to the Paradise framework. A usability study was conducted using a spelling system (self developed) versus the HeyAnita voice portal (professionally developed) to determine the usability of the systems using the SUM. The experiment was performed across three predefined metrics: user satisfaction, accuracy, and task completion time. Data was collected for all three metrics and analyzed with respect to the SUM. Initially, it was projected that the HeyAnita voice portal was a more usable system versus the spelling system. Therefore, it was not a surprise that the experiment results follow this same premise. However, the SUM was useful in identifying the extent to which the HeyAnita voice portal was more usable than the spelling system. Additionally, the SUM was successful in identifying areas where each system needs improvement. In summary, the SUM provides the designers with the ability to specify metrics that are critical to the specified speech application. Furthermore, the SUM allows the designers to specify weights associated with each metric. This gives the designers flexibility to customize the usability evaluation. The output of the SUM is a measure of the overall usability of the system obtained through experimentation using the designers’ metrics with the customized weights.

7. References


Ben Shneiderman, Clarifying Search A user-Interface Framework Text Searches, D-Lib Magazine, January 1997, ISSN 1082-9873


Miller, George A. The magical Number Seven, Plus or Minus Two: Limits on Our Capacity for Processing Information. Psychological Review Volume 63, 1956: 81-97.
